文章编号:1672-3961(2011)04-0020-04

一种快速 AP 聚类算法

刘晓勇[1,2,3],付辉[2]
(1. 中国科学院文献情报中心,北京 100190;
2. 广东技术师范学院计算机科学学院,广东 广州 510665;
3. 中国科学院研究生院,北京 100049)

摘要:Affinity propagation(AP)聚类算法中的一个重要参数 - 收敛系数(damping factor)对算法的运行效率有较大影响,而传统的 AP 算法中收敛系数常作为固定参数在算法运行中保持不变,因此 AP 算法的收敛性能对收敛系数初始值的选择比较敏感,针对这一问题提出了一种新的 AP 聚类算法:F-AP,该算法在传统 AP 聚类算法基础上引入收缩因子调节收敛系数,使其值能够随算法进程动态调整,以加速 AP 算法的收敛过程。在 3 个不同容量模拟数据集上进行了实验,结果表明,新算法能够有效加速收敛过程,并且能够保证与原算法相同的聚类结果;在标准数据集 Iris 上的聚类结果也表明了新算法具有较好的收敛性能。

关键词:聚类算法;吸引 - 传播聚类算法;收缩因子;振荡度
中图分类号:TP181 **文献标志码**:A

A fast affinity propagation clustering algorithm

LIU Xiao-yong[1,2,3], FU Hui[2]
(1. National Science Library, Chinese Academy of Sciences, Beijing 100190, China;
2. Department of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China;
3. Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: An important parameter of the affinity propagation algorithm (AP) damping factor, affects the speed of AP. Because the value of the damping factor is fixed in traditional AP algorithm, the convergence performance of the AP algorithm is sensitive to the parameter's choosing. A novel and fast AP algorithm, F-AP, was proposed. The new algorithm used the constriction factor to regulate damping factor dynamically. Three datasets and iris dataset were used to compare AP and F-AP. The numerical results showed that F-AP could effectively accelerate the convergence process.
Key words: clustering algorithm, affinity propagation, constriction factor, oscillation index

0 引言

吸引 - 传播聚类算法(affinity propagation clustering, AP)是由 B J Frey 和 D Dueck 于 2007 年提出的一种新的聚类算法[1],该算法无需事先定义类数,在迭代过程中不断搜索合适的聚类中心,自动从数据点间识别类中心(exemplars)的位置及个数。算法开始时把所有的数据点均视作类中心,通过数据点间的"信息传递"来实现聚类过程。与传统的 K 均值算法对初始类中心选择的敏感性相比,AP 算法是一种确定性的聚类算法,多次独立运行的聚类结果一般都十分稳定。该算法以其简单、高效的优势已广泛应用于多种领域,如:设施选址[2-4]、图像识别[5],图像分割[6-7]、文本挖掘[8],生物医学[1,9],视频关键帧提取[10]和图像检索[11]等方面。国内的王开军,谢信喜、肖宇、谷瑞军、董俊及李雅芹等人针对AP算法的不足提出了多种改进方法,也取得了

收稿日期:2011-02-14
基金项目:广东高校优秀青年创新人才培养计划项目(LYM10097);2011 年广东技术师范学院科研项目(自然科学)
作者简介:刘晓勇(1979 -),男,河南信阳人,讲师,主要研究方向为智能优化算法,文本挖掘等. E-mail: lxyong420@126.com

较好的效果[12-19]。

本文首先介绍了吸引 - 传播聚类算法,然后根据该算法的不足提出基于伸缩因子的吸引传播聚类算法 F-AP,在数值实验部分先分析了 AP 算法的重要参数之一(收敛系数 λ)对聚类结果的影响,然后在 3 个实验数据集上将 F-AP 与原 AP 算法进行了性能比较。

1　基于收缩因子的 AP 算法

1.1　AP 算法

AP 算法[1]不需要数据集具有某种特殊的结构,主要根据 N 个样本点之间的相似度进行聚类,这些相似度组成 $N \times N$ 的相似度矩阵 $\boldsymbol{S}$,如:$\boldsymbol{S}(i,j)$ 表示样本点 i 和样本点 j 之间的相似度。矩阵 $\boldsymbol{S}$ 主对角线上的数值又称为 Preference,该值是对应的样本点能否成为聚类中心的评判标准,一般来说,其值越大表示这个点成为聚类中心的可能性就越大。AP 算法主要依靠一种"消息传递"机制实现数据集的聚类。这种消息传递机制中主要包含两类信息:Responsibility 和 Availability。Responsibility 表示样本点对不同的候选类中心发出的信息,表明候选类中心相应于该样本点作为潜在类中心的适合程度,该值越大表明候选类中心越可能成为实际的类中心;Availability 表示候选类中心对样本点发出的信息,表明该样本点相应于候选类中心的聚合程度,该值越大表明样本点越可能属于某一类。AP 算法通过迭代过程不断更新每一个点的 Responsibility 和 Availability 值,直到自动产生若干个类中心,同时将其余的数据点分配到相应的类团中。AP 算法的步骤如下:

Step 1　算法初始化,计算初始相似度矩阵 $\boldsymbol{S}$;对 P 赋初值。

Step 2　计算样本点间的 Responsibility 值

$$R(i,k) \leftarrow s(i,k) - \max_{j \neq k}(s(i,j) + A(i,j)), \quad (1)$$

$A(i,j)$ 表示 j 对于 i 的 Availability 值。

Step 3　计算样本点间的 Availability 值。

$$A(i,k) \leftarrow \min\{0, R(k,k) + \sum_{j \neq i,k} \max(0, R(j,k))\}, \quad (2)$$

$$A(k,k) \leftarrow \sum_{j \neq k} \max(0, R(j,k))。 \quad (3)$$

Step 4　Responsibility 和 Availability 的更新:

$$R_{i+1}(i,k) = \lambda \cdot R_i(i,k) + (1-\lambda) \cdot R_{i+1}^{\text{old}}(i,k), \quad \lambda \in [0.5,1), \quad (4)$$

$$A_{i+1}(i,k) = \lambda \cdot A_i(i,k) + (1-\lambda) \cdot A_{i+1}^{\text{old}}(i,k), \quad \lambda \in [0.5,1), \quad (5)$$

λ 是收敛系数,主要用于调节算法的收敛速度及迭代过程的稳定性。

$$A_{i+1}(k,k) = P(k) - \max[A_{i+1}(k,j) + S_{i+1}(k,j)], \quad j \in \{1,2,3,\cdots,N\}, j \neq k。 \quad (6)$$

Step 5　如果迭代次数超过设定的最大值或者当聚类中心在若干次迭代中不发生改变时终止计算,确定类中心及各类的样本点;否则返回 Step2,继续计算。

1.2　收缩因子

Clerc Maurice 的研究表明使用收缩因子(constriction factor)可以有效地保证算法收敛[9]。收缩因子的定义公式为

$$\rho = \frac{2}{|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}|}, \varphi > 4。 \quad (7)$$

在本研究提出的快速 AP 算法(F-AP)中,为了加速算法收敛,对 Responsibility 和 Availability 的更新公式作出如下变化:

$$R_{i+1}(i,k) = \rho \cdot \lambda \cdot R_i(i,k) + (1-\lambda) \cdot R_{i+1}^{\text{old}}(i,k) \quad (8)$$

$$A_{i+1}(i,k) = \rho \cdot \lambda \cdot A_i(i,k) + (1-\lambda) \cdot A_{i+1}^{\text{old}}(i,k), \lambda \in [0.5,1)。 \quad (9)$$

在数值实验部分,φ 取值为 4.1,因此 ρ = 0.729。

2　数值实验

为了验证新算法的性能,本研究进行了两组实验,第 1 组实验主要考查收敛系数 λ 对聚类结果的影响;第 2 组实验主要比较 AP 与 F-AP 的性能。

2.1　收敛系数 λ 对聚类结果的影响

为了说明收敛系数 λ 对聚类结果的影响,模拟了一个有 2 000 个样本点的数据集,该数据集的数据点在[0,1]上随机生成,然后以 AP 算法为基础,针对 λ 取不同数值分析聚类结果。为了反映算法的稳定程度定义了一个振荡度指数,该值越小说明算法在迭代过程中振荡越小,算法运行越平稳。

定义　振荡度(oscillation index,OI)

$$\text{OI} = \frac{\text{count}[(e_{i+1} - e_i) < 0]}{T}, \quad i = 1,2,3,\cdots N。 \quad (10)$$

这里,e_i 表示第 i 次迭代时样本点间相似度的值。T 表示算法开始稳定时已经迭代的次数。

数值实验中 λ 分别取 0.7,0.8 和 0.9,表 1 说明了当 λ 取不同值时的数值比较结果,图 1 是 λ

3 种不同取值时的收敛曲线。从实验结果中可以发现,当 λ 越大时算法消除振荡的效果越好,迭代曲线越平稳。其中,λ =0.9 时振荡度最小,而 λ =0.7 时,振荡度最大。但从迭代曲线中可以发现,当 λ 值越大时,算法的收敛速度就越慢,其中,当 λ =0.9 时需要迭代 231 次才结束,而当 λ 分别取 0.7 和 0.8 时,迭代次数均不超过 100 次。因此本组实验表明,一个合适的收敛系数值对 AP 聚类算法在振荡度和收敛速度方面有较明显的影响。

表 1 数值比较结果(data 2000)

Table 1 The comparison among different values of λ (data 2000)

λ	类数	迭代次数	样本点间相似度值	振荡度
0.7	39	94	-18.726	0.178
0.8	40	86	-18.482	0.108
0.9	39	231	-18.442	0

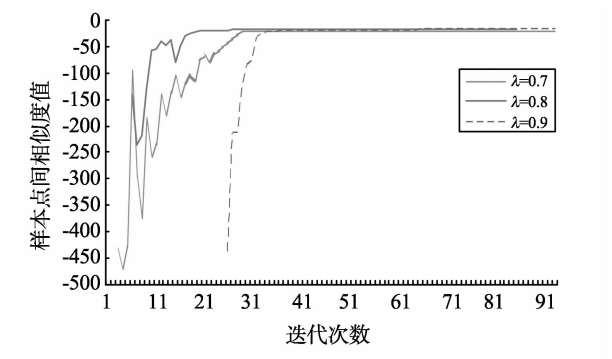

图 1 收敛曲线

Fig. 1 Convergence curves

2.2 F-AP 与 AP 性能比较

为了比较 F-AP 与 AP 聚类算法的性能,这里采用3 个模拟数据集,分别随机生成 100、500、1 000 个在[0,1]上均匀分布的数据点。表 2 是 2 个算法在不同数据集上的运行结果。图 2、图 3、图 4 是两种算法对 3 个数据集的聚类图,从中可以发现 2 种算法聚类的结果完全一致,这说明 F-AP 具有与 AP 一样的聚类性能。

表 2 F-AP 与 AP 的数值比较结果

Table 2 The comparison between F-AP and AP

数据集	算法	聚类个数	运行时间/s	迭代次数	数据点间相似度值
Data 100	AP	8	2.641	101	-3.974 44
	F-AP	8	2.454	93	-3.974 44
Data 500	AP	20	5.937	100	-9.526 21
	F-AP	20	5.344	93	-9.526 21
Data 1000	AP	27	15.968	110	-13.256 2
	F-AP	27	14.11	96	-13.256 2

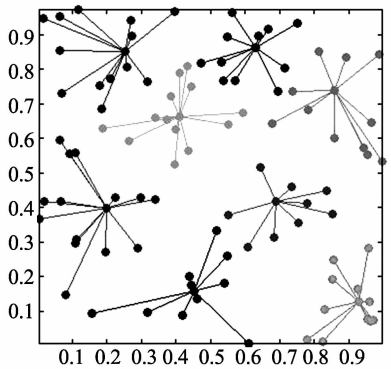

图 2 AP 和 F-AP-聚类图(data 100)

Fig. 2 Clustering diagram of AP and F-AP (data 100)

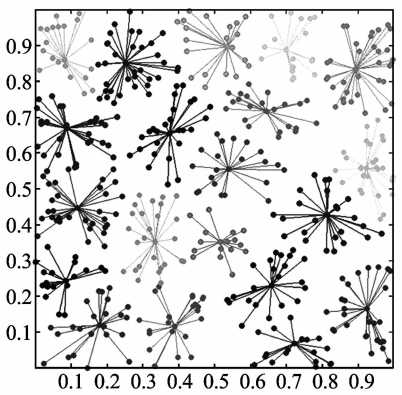

图 3 AP 和 F-AP-聚类图(data 500)

Fig. 3 Clustering diagram of AP and F-AP (data 500)

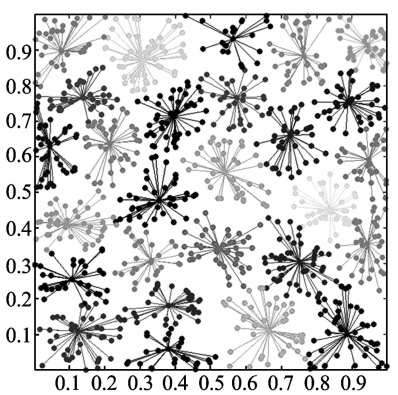

图 4 AP 和 F-AP-聚类图(data 1000)

Fig. 4 Clustering diagram of AP and F-AP (data 1000)

从上述 3 个数据集上的数值实验可以发现,F-AP 算法仅需要较少的迭代次数和运行时间就可以达到与 AP 算法一样的聚类结果。同时,在数据集 data 100,data 500 和 data 1000 中,F-AP 与 AP 相比在运行时间上分别提升了 7.08%,9.99% 和 11.6%,这说明随着数据集容量的增加,F-AP 的运行效率优势更加明显。

为了进一步对比两种算法的性能,本文使用聚类分析算法中常用的 Iris 数据集来进行算法比较。Iris 是一种鸢尾花数据集,包含 150 个 4 维数据点,共分为 3 大类,每类各包含 50 个数据点,其中第 1

类较为明显,而第 2、3 类较不明显。两种算法在该数据集运行后,F-AP 算法需要迭代 63 次收敛;而 AP 算法则需要迭代 67 次才达到收敛。

3　结语

AP 聚类算法以其高效的性能已经得到了比较广泛的应用,在分析了 AP 算法中的一个重要参数 λ 对聚类结果影响的基础上,将收缩因子引入到 AP 算法中,提出一种快速的聚类算法:F-AP,在 3 个数据集上的数值实验表明,F-AP 算法经过较少的迭代次数和较少的运行时间就能得到与 AP 算法一样的聚类性能,并且数据集越大这种速度优势就愈明显。同时在常用数据集 Iris 上的算法执行效果也表明提出的新算法具有较好的收敛性能。

参考文献:

[1] FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315 (5814): 972-976.

[2] LAZIC N, GIVONI Inmar E, AARABI Parham, et al. FLoSS: Facility location for subspace segmentation[C]// Proceedings of 12th International Conference on Computer Vision (ICCV). Kyoto: IEEE Press, 2009: 825-832.

[3] LAZIC Nevena, FREY Brendan J, AARABI Parham. Solving the uncapacitated facility location problem using message passing algorithms[C]// Proceedings of 13th International Conference on Artificial Intelligence and Statistics (AISTATS). Sardinia: Microtome Publishing, 2010: 429-436.

[4] 唐东明, 朱清新, 杨凡,等. 基于仿射传播聚类的大规模选址布局问题求解[J]. 计算机应用研究, 2010,27(3): 841-844.
TANG Dongming, ZHU Qingxin, YANG Fan, et al. Solving large scale location problem using affinity propagation clustering[J]. Application Research of Computers, 2010, 27(3): 841-844.

[5] DUECK D, FREY B J. Non-metric affinity propagation for unsupervised image categorization[C]// Proceedings of 11th International Conference on Computer Vision (ICCV). Rio de Janeiro: IEEE Press, 2007: 1-8.

[6] 张仁彦, 赵洪亮, 卢晓, 等. 基于相似性传播聚类的灰度图像分割[J]. 海军工程大学学报, 2009,21(3):33-37.
ZHANG Renyan, ZHAO Hongliang, LU Xiao, et al. Grey image segmentation method based on affinity propagation clustering [J]. Journal of Naval University of Engineering, 2009, 21(3): 33-37.

[7] GIVONI I E, FREY B J. Semi-supervised affinity propagation with instance-level constraints [C]//Proceedings of 12th International Conference on Artificial Intelligence and Statistics (AISTATS). Florida: Microtome Publishing, 2009: 161-168.

[8] 管仁初, 裴志利, 时小虎, 等. 权吸引子传播算法及其在文本聚类中的应用[J]. 计算机研究与发展, 2010, 47(10): 1733-1740.
GUAN Renchu, PEI Zhili, SHI Xiaohu, et al. Weight affinity progagation and its application to text clustering [J]. Journal of Computer Research and Development, 2010, 47(10):1733-1740.

[9] DUECK D, FREY B J, JOJIC N, et al. Constructing treatment portfolios using affinity propagation [C]// Proceedings of International Conference on Research in Computational Molecular Biology (RECOMB). Singapore: Springer, 2008: 360- 371.

[10] 许文竹, 徐立鸿. 基于仿射传播聚类的自适应关键帧提取 [J]. 计算机科学, 2010(1): 268-270.
XU Wenzhu, XU Lihong. Adaptive key-frame extraction based on affinity propagation clustering[J]. Computer Science, 2010(1):268-270.

[11] 向培素. 一种基于近邻半监督聚类算法的图像检索系统研究[J]. 西南民族大学学报: 自然科学版, 2010, 36(4):624-627.
XIANG Peisu. New CBIR system based on the af finity propagation clustering algorithm[J]. Journal of Southwest University for Nationalities: Natural Science Edition, 2010, 36(4): 624-627.

[12] 王开军, 张军英, 李丹, 等. 自适应仿射传播聚类[J]. 自动化学报, 2007, 33(12):1242-1246.
WANG KaiJun, ZHANG Junying, LI Dan, et al . Adaptive affinity propagation clustering [J]. Acta Automatica Sinica, 2007, 33 (12):1242-1246.

[13] 王开军, 李健, 张军英, 等. 半监督的仿射传播聚类[J]. 计算机工程, 2007, 33(23): 197-198,201.
WANG Kaijun, LI Jian, ZHANG Junying, et al. Semi-supervised affinity propagation clustering [J]. Computer Engineering, 2007, 33(23): 197-198,201.

[14] 王开军, 郑捷. 指定类数下仿射传播聚类的快速算法[J]. 计算机系统应用,2010,19(7):207-209.
WANG KaiJun, ZHENG Jie. Fast algorithm of affinity propagation clustering under given number of clusters [J]. Computer Systems and Applications. 2010, 19(7): 207-209.

[15] 谢信喜,王士同. 适用于区间数据的基于相互距离的相似性传播聚类[J]. 计算机应用,2008, 28(6): 1441-1443.
XIE Xinxi, WANG Shitong. Affinity propagation clustering for symbolic interval data based on mutual distance [J]. Computer Application, 2008, 28(6):1441-1443.

[16] 肖宇, 于剑. 基于近邻传播算法的半监督聚类 [J].

软件学报, 2008, 19(11): 2803- 2813.
XIAO Yu, YU Jian. Semi-supervised clustering based on affinity propagation algorithm[J]. Journal of Software, 2008, 19(11):2803-2813.

[17] 谷瑞军, 汪加才, 陈耿, 等. 面向大规模数据集的近邻传播聚类[J]. 计算机工程,2010,36(23):22-24.
GU Ruijun, WANG Jiacai, CHEN Geng, et al. Affinity propagation clustering for large scale dataset[J]. Computer Engineering, 2010, 36(23): 22-24.

[18] 董俊, 王锁萍, 熊范纶. 可变相似性度量的近邻传播聚类[J]. 电子与信息学报, 2010,32(3):509-514.
DONG Jun, WANG Suoping, XIONG Fanlun. Affinity propagation clustering based on variable-similarity measure[J]. Journal of Electronics & Information Technology, 2010, 32(3):509-514.

[19] 李雅芹, 杨慧中. 基于仿射传播聚类和高斯过程的多模型建模方法[J]. 计算机与应用化学, 2010,27(1): 51-54.
LI Yaqin, YANG Huizhong. Multi-model modeling method based on affinity propagation clustering and Gaussian processes[J]. Computers and Applied Chemistry, 2010, 27(1):51-54.

[20] CLERC Maurice. The swarm & the queen towards a deterministic and adaptive particle swarm optimization [C]// Proceedings of Congress on Evolutionary Computation, Washington: IEEE Press, 1999: 1951-1957.

(编辑:孙培芹)